Environmental Health Perspectives
Vol. 32, pp. 181-192, 1979

Statistical Methods for Censored Survival Data

by Norman Breslow*

Methods of statistical analysis of censored survival times are briefly reviewed and illustrated by application to clinical trials data. These include estimation of the survival curve, nonparametric tests to compare several survival curves, tests for trend, and regression analysis. Extensions of the methodology are made for application to epidemiologic case-control studies. These are used to estimate relative risks for leukemia associated with radiation exposures. A final section provides some annotated references to the recent literature.

Introduction

Censored survival data arise in a wide variety of statistical investigations. In clinical trials one measures duration of response from start of treatment until relapse or death due to disease. Observations on response time are censored for those subjects still in remission at the study's end, as they are for patients lost to followup during the course of the study. Animal carcinogenesis studies, such as used by the United States Food and Drug Administration to determine the safety of food additives, provide another example. Here the endpoint is the age at diagnosis of a particular kind of cancer, censorship being imposed by death due to other causes, natural or artificial. In tests of the reliability of airplane components, failure times are measured from the start of testing until failure of the component, with censorship imposed by the failure of other components or the necessity of analyzing the data before all items have failed.

Figure 1 illustrates the results for the control group in a clinical trial designed to investigate the effects of combined chemotherapy as an adjunct to surgery and radiation in the treatment of childhood rhabdomyosarcoma (*1*). The endpoint for analysis was the reappearance of tumor, whether at the site of original treatment or through distant metastasis. Children who remained disease-free at the time the data were analyzed had censored observations. In addition to the control arm IA, there were two groups of children who received the drugs actinomycin-D (AMD) and vincristine (VCR): group IB patients were concurrently randomized with the controls, both these groups having apparently had their tumors completely resected; group IIA consisted of patients with microscopic residual disease at the margin of surgical resection.

Interim data from all three arms are presented in Table 1. Note that the censored observations for arm IA, those in the column labeled "disease-free", are smaller in the table than they are in the figure. This is because the figure was drawn from data computed at a later point in time, when additional follow-up was available for patients who had not already died.

Analysis of such data has several goals. For each of the comparison groups one wants an estimate of the survival curve, the probability of surviving t units of time. Statistical tests are required to determine whether the observed differences between the curves are real, or are simply chance effects. If real, a method of quantifying the nature of the differences is desirable. Finally there may be available concomitant observations, including continuous measurements such as age at diagnosis, whose joint effects on survival are important to determine.

Estimation of Survival Curves

When analyzing several groups of survival times the first step is to form a series of $2 \times r$ tables as

*Department of Biostatistics SC-32, University of Washington, Seattle, Washington 98195.

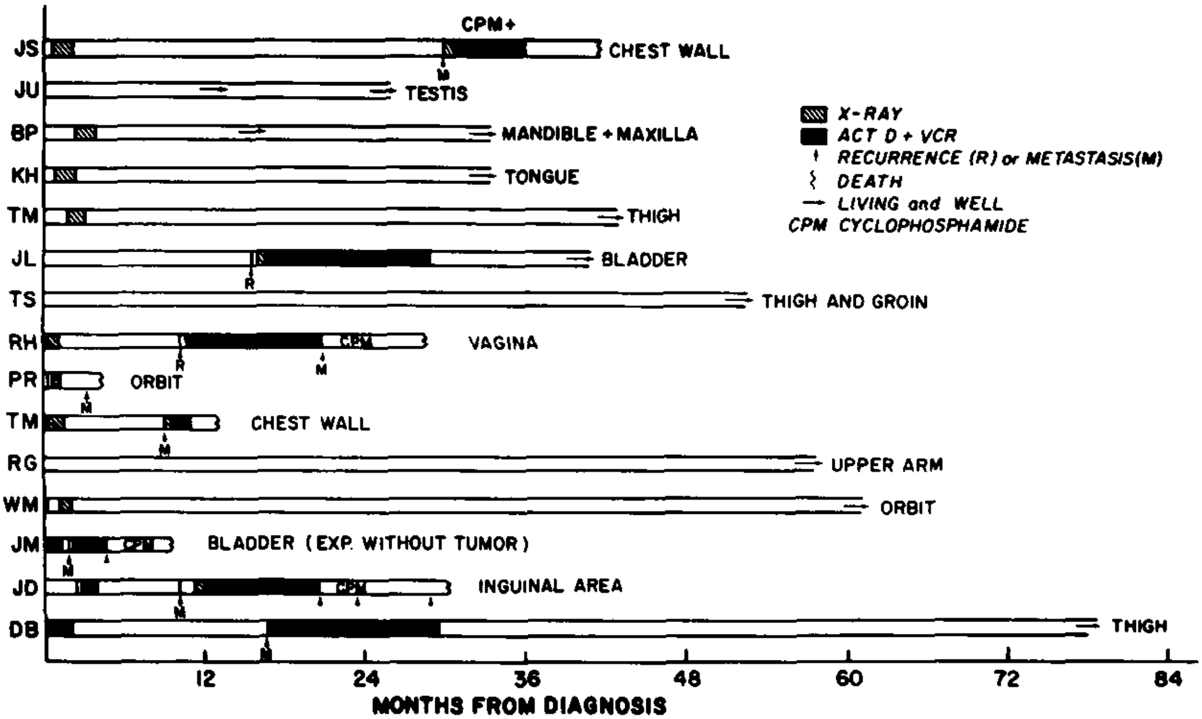

FIGURE 1. Course of patients following surgery for rhabdomyosarcoma in Part IA, control group.

shown in Table 2. One table is formed for each of the K distinct times $t_1 < t_2 < \ldots < t_K$ at which deaths (or failures or relapses) occur. The column totals n_{ik} refer to the total number of subjects in the i^{th} group who remain "at risk", i.e., alive and under observation, just prior to time t_k. The tabular entries d_{ik} and s_{ik} denote the numbers of those who die at t_k, and survive t_k, respectively. Table 3 illustrates the calculation of the first three such tables for the data in Table 1. Here $r = 3$ and $t_1 = 2$, $t_2 = 3$, and $t_3 = 9$ months. Note that the tables for increasing t_k refer to a constantly diminishing population "at risk" as additional subjects die or are withdrawn (censored) from further observation.

Table 1. Interim clinical trial data: time from start of treatment to relapse or last observation for three treatment groups.

Time, months					
Tumor completely resected				Microscopic residual	
Surg + x-ray (IA)		Surg + x-ray + AMD + VCR (IB)		Surg + x-ray + AMD + VCR (IIA)	
Relapsed	Disease-free	Relapsed	Disease-free	Relapsed	Disease-free
2	12	9	12	37	25
3	15	16	19		28
9	18	19	20		29
10	24		20		38
10	36		24		42
15	40		24		45
16	45		30		47
30			31		48
			34		50
			42		52
			44		
			53		
			59		
			62		

Kaplan and Meier (2) derived the maximum likelihood nonparametric estimate of the survival curve based on censored data. This may be calculated recursively, starting from $\hat{P}(t_0) = 1$, and by using the formula (1):

$$\hat{P}(t_k) = \hat{P}(t_{k-1})\,(s_k/n_k) \qquad (1)$$

Table 2. Formation of 2 × r contingency tables comparing death rates among r treatment groups at each distinct time of death.

Patients followed to time t_k for various treatment groups					
	1	2	—	r	Totals
Deaths (at t_k)	d_{1k}	d_{2k}	—	d_{rk}	D_k
Survivors	s_{1k}	s_{2k}	—	s_{rk}	S_k
Total "at risk"	n_{1k}	n_{2k}	—	n_{rk}	N_k

Table 3. Illustration of 2 × r tables for data in Table 1.

t, months		IA	IB	IIA	Total
2	Relapsed	1	0	0	1
	Disease-free	14	17	11	42
	"At risk"	15	17	11	43
	Expectations:	0.349	0.395	0.256	1.000
3	Relapsed	1	0	0	1
	Disease-free	13	17	11	41
	"At risk"	14	17	11	42
	Expectations:	0.333	0.405	0.262	1.000
9	Relapsed	1	1	0	2
	Disease-free	12	16	11	39
	"At risk"	13	17	11	41
	Expectations:	0.634	0.829	0.537	2.000

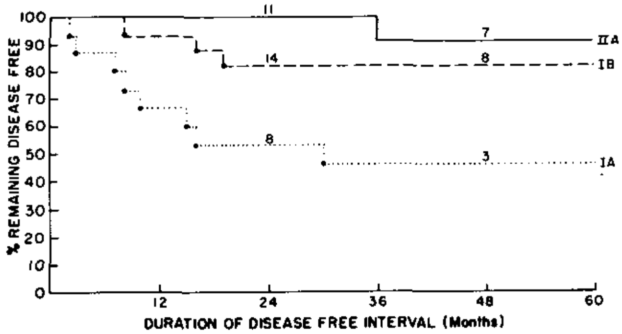

FIGURE 2. Duration of the disease-free interval in patients from Part IA (control), IB (treated), and IIA (microscopic residual, treated). Shown above each curve at 24 and 48 months are the numbers of patients known to be disease free after those time periods.

for $k = 1, 2, \ldots, K$. (The group index i has been suppressed for clarity.) In other words, the probability of surviving past t_k is estimated as the probability of surviving past t_{k-1} times the conditional probability of surviving past t_k given survival to t_k. Because of this multiplicative structure, Kaplan and Meier refer to their estimate as the product limit (PL) estimate. In case there is no censorship in the data, it reduces to the familiar empirical distribution function.

Table 4 shows the calculation of the relapse-free survival curve from the interim data in Table 1 for treatment group IA. The corresponding curves calculated from final study data for all three treatment curves are shown in Figure 2. Numbers above each curve at annual intervals in this figure refer to numbers of patients still at risk in each group. These are an important means of judging the stability of the estimates, which can in fact be quite unstable in the "tail" of the survival distribution where few subjects remain at risk.

The variance of the PL estimate may also be calculated recursively, starting from $\hat{V}\{P(t_0)\} = 0$. One uses the formula (2)

$$\hat{V}\{P(t_k)\} = \hat{V}\{P(t_{k-1})\}(s^2_k/n^2_k) + \{\hat{P}(t_k)\}^2 [d_k/(n_k s_k)] \quad (2)$$

with the understanding that it is applied sequentially to tied observations. In large samples, $\hat{P}(t)$ is approximately normally distributed with mean equal to the true survival function $P(t)$ and a variance estimated as shown above (*2*, *3*). Note that neither $\hat{P}(t)$ nor $\hat{V}\{P(t)\}$ will change after the last uncensored response time in each group, even though additional subjects continue to be withdrawn from observation. In this region the estimated variance often does not accurately reflect the true variability in the survival curve, which will be substantial unless large numbers of subjects remain on study.

Table 4. Estimation of survival curve for group IA.

Time, months t_k	Number at risk n_{1k}	Number surviving s_{1k}	Conditional probability $\hat{p}(t_k)$	Survival probability $\hat{P}(t_k)$
2	15	14	0.933	0.933
3	14	13	0.929	0.866
9	13	12	0.923	0.799
10	12	10	0.833	0.666
15	9	8	0.888	0.592
16	7	6	0.857	0.507
30	4	3	0.750	0.381

Comparison of Survival Curves

A simple but powerful non-parametric test for the comparison of r survival curves with censored data may also be calculated from the basic data shown in Table 2. This test exploits the fact that, under the null hypothesis of no difference in the underlying survival distributions and conditional upon fixed values for the marginal totals in each 2 × r table, the vector $\mathbf{d}_k = (d_{1k}, \ldots, d_{rk})'$ of observed deaths at t_k has an r-dimensional hypergeometric distribution. Consequently the null expectation of the number of deaths in group i at t_k is

$$e_{ik} = E(d_{ik}) = n_{ik}(D_k/N_k) \quad (3)$$

i.e., the number at risk in the i-th group times the death rate for all r groups combined (see Table 3 for an illustration of this calculation). The covariance matrix $\mathbf{V}_k$ of $\mathbf{d}_k$ has, under the null hypothesis, an (i,j) component equal to

$$\|\mathbf{V}_k\|_{ij} = \begin{cases} \dfrac{n_{ik}(N_k - n_{ik})\, D_k S_k}{N^2{}_k\,(N_k - 1)} & i = j \\ -\dfrac{n_{ik} n_{jk} D_k S_k}{N^2{}_k\,(N_{k-1})} & i \neq j \end{cases} \qquad (4)$$

The main idea behind the test is to sum up the statistics calulated from each of the K tables into a vector $\mathbf{0} = \Sigma_k \mathbf{d}_k$ of observed numbers of deaths in each group, a vector $\mathbf{E} = \Sigma_k \mathbf{e}_k$ of expected numbers of deaths, and a summary covariance matrix $\mathbf{V} = \Sigma_k \mathbf{V}_k$. Since the $2 \times r$ tables refer to overlapping sets of subjects they are not, strictly speaking, statistically independent. Nevertheless Cox (*4*) has shown that $\mathbf{V}$ is an appropriate large sample covariance matrix for $\mathbf{O}$-$\mathbf{E}$. Since $\Sigma 0_i = \Sigma E_i$, i.e. the totals of observed and expected deaths in all r groups agree, $\mathbf{V}$ is singular. However, defining $\mathbf{O}^*$ and $\mathbf{E}^*$ to be the first $r - 1$ components of $\mathbf{O}$ and $\mathbf{E}$, and $\mathbf{V}^*$ to be the $(r - 1) \times (r - 1)$ upper left hand corner of $\mathbf{V}$, a test statistic for testing equality of the r survival curves is obtained as

$$T_1 = (\mathbf{O}^* - \mathbf{E}^*)' \mathbf{V}^{*-1} (\mathbf{O}^* - \mathbf{E}^*) \qquad (5)$$

This is approximately distributed as chi-square on $r - 1$ degrees of freedom under the null hypothesis.

The test T_1 was first proposed for survival data by Mantel (*5*). Cox (*6*) later derived it from likelihood theory under the proportional hazards (PH) model, in which the instantaneous death rates in the r groups are assumed to be in constant ratio throughout the follow-up period (see below). Peto and Peto (*7*) argued that it was an asymptotically efficient test under Cox's model and named it the "log rank" test.

A conservative approximation to T_1 which requires no matrix inversion is given by the familiar chi-square formula

$$T_2 = \sum_{i=1}^{r} (O_i - E_i)^2 / E_i \qquad (6)$$

While $T_2 \leq T_1$, in fact the two statistics will be quite close, provided that there are few ties among the uncensored survival times (i.e., most of the D_k in Table 2 are unity) and that the patterns of censorship operating in the r groups are not grossly different (*8*, *9*). Note that the ½ continuity correction should not be used with survival data.

Table 5 illustrates the manner of presentation of the summary and test statistics. Note the calculation of the ratio O/E of observed to expected numbers of deaths in each treatment group. These are very useful as measures of treatment effect, since their ratios, e.g., $O_1/E_1 \div O_2/E_2$, approximate the ratios of death rates in the respective treatment groups (*10*).

Table 5. Summary statistics for interim clinical trial data.

	Treatment group		
	IA	IB	IIA
No. of patients (N)	15	17	11
Relapses observed (O)	8	3	1
Relapses expected (E)	3.11	4.99	3.90
O/E	2.57	0.60	0.26
W scores	−179	68	111
Covariance matrix $\mathbf{V}$			
	2.22		
	−1.27	2.86	
	−0.94	−1.59	2.53
Covariance matrix $\mathbf{V}_w$			
	2954		
	−1748	3572	
	−1206	−1824	3030

[a]Test statistics:

$T_1 = 11.10$, 2 df, $p = 0.004$

$T_2 = 10.77$, 2 df, $p = 0.005$

$T_3 = 11.41$, 2 df, $p = 0.003$

$T_4 = 11.20$, 2 df, $p = 0.004$

Alternate Weighting Schemes

The summary statistics $O - E$ weight the observed differences $d_k - e_k$ in each table in a manner which is appropriate to the PH model already mentioned. However this is not the only possible weighting scheme. Multiplying the observed differences by N_k, the total number of subjects in the k-th table, and then summing, gives more weight to the earlier times t_k when larger numbers are at risk. This leads to the scores

$$W_i = \sum_{k=1}^{K} \{N_k d_{ik} - n_{ik} D_k\} \qquad (7)$$

covariance matrix

$$\mathbf{V}_w = \sum_{k=1}^{K} N_k{}^2 \mathbf{V}_k \qquad (8)$$

and test statistic

$$T_3 = \mathbf{W}^{*\prime} \mathbf{V}_w{}^{*-1} \mathbf{W}^* \qquad (9)$$

where asterisks (*) denote the corresponding $r - 1$ dimensional quantities. A conservative approximation to T_3 not requiring matrix inversion is

$$T_4 = \sum_{i=1}^{r} W_i^2/G_i \tag{10}$$

with

$$G_i = \sum_{k=1}^{K} \{N_k D_k S_k n_{ik}/N_{k-1}\} \tag{11}$$

The scores W_i may also be obtained from a pairwise comparison of the observations in the i-th treatment group with those in the remaining $r - 1$ groups. Each such pair is assigned the value +1 (or −1) according as the true survival time for the first pair member is known to be smaller than (or larger than) that for the second member. Ties or indeterminate comparisons due to censorship are assigned 0 values. Gehan (*11*) suggested the use of such scores for the comparison of two samples ($r = 2$). In this case T_4 reduces to the familiar Wilcoxon rank sum test in the absence of ties and censorship. Breslow (*12*) extended this work to $r > 2$ samples, proposing the covariance matrix $\mathbf{V}_w$ and the statistic T_3. These latter statistics, like **V** and T_1, are valid for situations where the patterns of censorship operative in the r treatment groups are unequal, as in animal carcinogenesis studies where there is differential toxic mortality. The conservative approximation T_4, like T_2, is strictly valid only where there is equality of censorship.

In practice, the tests T_1 and T_3 often yield rather similar numerical values (see Table 5). However, this is not always true, and some comment on the proper interpretation when only one statistic is significant is in order. Since T_3 weights early values more heavily, it may achieve significance when there is an early separation between the survival curves which later come together or even cross over. T_1 gives more weight to the later appearing deaths. A large discrepancy between T_1 and T_3 generally indicates an interaction between treatment and time on the instantaneous death rates, which is worthy of investigation in its own right.

Testing for Trend

Often the r comparison groups correspond to r levels $x_1 < x_2 < \dots x_r$ of a quantitative variable such as dose. Global chi-square tests such as T_1 through T_4 lack statistical power in such situations since they take no account of the natural order of the groups. One needs a single degree of freedom test for trend in survival with increasing dose.

Fortunately such a test is readily calculated from the summary statistics already at hand. In the case of the log rank analysis

$$T_5 = \frac{\{\mathbf{x}'\,(\mathbf{O} - \mathbf{E})\}^2}{\mathbf{x}'\mathbf{V}\mathbf{x}} \tag{12}$$

is a single degree of freedom chi-square for a linear trend of O-E with x. Tarone (*13*) has suggested using $T_6 = T_1 - T_5$ as a chi-square on r-2 degrees of freedom for deviations from linearity. An approximation to T_5 which only requires calculation of the **O** and **E** vectors is given by

$$T_7 = \frac{\{\Sigma x_i\,(0_i - E_i)\}^2}{\Sigma x^2{}_i E_i - (\Sigma x_i E_i)^2/\Sigma E_i} \tag{13}$$

Similarly, when using the W scores,

$$T_8 = \frac{\{\mathbf{x}'\mathbf{W}\}^2}{\mathbf{x}'\mathbf{V}_w\mathbf{x}} \tag{14}$$

provides a test for linear trend and $T_9 = T_3 - T_7$ a test for deviations from linearity.

To illustrate these calculations by using the summary data in Table 5, make the fictitious assumption that the three treatment groups IA, IB, and IIA correspond to dose levels $x_1 = 0$, $x_2 = 1$, $x_3 = 2$. The statistics for trend are $T_5 = 9.17$ ($p = 0.002$), $T_7 = 8.72$ ($p = 0.003$), and $T_8 = 10.02$ ($p = 0.002$). The deviation chi squares are $T_6 = 1.93$ (NS) and $T_9 = 1.39$ (NS). Hence, this would be a case where the already significant differences are largely explained on the basis of an apparent linear trend in survival with increasing dose.

Adjustment by Stratification

When the r comparison groups differ with respect to factors which influence survival, an analysis which corrects for their possible confounding effects is needed. This may be carried out very simply by dividing the population into strata which are more or less homogeneous internally with respect to the confounding factors. (Of course the number of confounders which may be accommodated simultaneously in this fashion is limited, since if strata become very large in number and small in size a substantial loss of comparative information may result.) Separate survival analyses are performed within each stratum by calculating the summary statistics **O, E,** and **V** defined earlier. These are cumulated by simple addition over strata and used to calculate adjusted test statistics T_1, T_2, T_5, T_6 and T_7, in which the cumulated summary statistics replace the stratum specific ones. Likewise adjusted versions of T_3, T_8, and T_9 use the cumulated **W** and $\mathbf{V}_w$ statistics.

Such a stratified analysis was used for a trial of maintenance chemotherapy for children with acute lymphocytic leukemia (*14*). For this disease it is well known that the diagnostic white blood count (WBC) is an important prognostic factor. The treatment group, consisting of 152 children who received actinomycin (AMD) in addition to standard maintenance drugs, had a median WBC of 10,067; whereas the control group, consisting of 116 children who did not receive AMD, had a median WBC of 14,280. An analysis ignoring this difference in WBC compared the observed number of relapses in the treated and control groups, $O_1 = 100$ and $O_2 = 81$, with expected numbers of $E_1 = 113.00$ and $E_2 = 68.00$. This yielded an (unadjusted) chi-square of $T_2 = 3.98$, which is just on the borderline of 5% statistical significance.

In order to determine whether the apparent effectiveness of AMD was due, at least in part, to the generally lower WBC's among treated patients, the entire sample of 268 children was divided into three strata as shown in Table 6. A separate calculation of the observed and expected numbers of relapses was made within each stratum, so the totals of O's and E's taken across each row of Table 6 agree. The adjusted expected numbers, $E_1 = 110.96$ and $E_2 = 70.04$, are now closer to the observed numbers and give an adjusted chi-square of $T_2 = 2.80$, which is no longer statistically significant.

The estimated ratio of relapse rates in the treated vs. control group is 100/113.0 ÷ 81/68.0 = 0.74. When adjusted for WBC in three strata, it is 100/110.96 ÷ 81/70.04 = 0.78, again indicating less of a difference between treatment and control group.

Regression Analysis of Survival Data

If the number of confounding variables is large, stratification breaks down since there will be many strata with just one or a few subjects. It may also be of interest to quantify the relationship between survival and several discrete or continuous and concomitant variables. This situation calls for a regression approach.

The usual regression model specifies that the survival times, or some transform such as their logarithm, are equal to a linear combination of the concomitant variables plus a random error term. Unfortunately, to generalize such models for use with censored data is awkward and computationally involved. Thus considerable interest was aroused by Cox (*6*) when he proposed a model formulated in terms of the effect of the regression variables on instantaneous death rates rather than on times of death *per se*. This model turned out to be quite tractable computationally and, as an added benefit, avoided any parametric assumptions about the shape of the underlying survival curve.

Cox's model is defined in terms of the time t specific death rate or hazard function $\lambda(t|\mathbf{z})$ for an individual having a p-vector of covariates $\mathbf{z}$. Specifically he assumes $\lambda(t|\mathbf{z}) = \exp(\beta'\mathbf{z})\lambda_0(t)$, where β is an unknown p-vector of parameters (regression coefficients), while $\lambda_0(t)$ is the unknown hazard or death rate function for an individual with a standard ($\mathbf{z} = \mathbf{O}$) set of convariates. A consequence of this model is that the ratio of hazard functions for two individuals with different sets of covariates,

$$\frac{\lambda(t|\mathbf{z}_1)}{\lambda(t|\mathbf{z}_2)} = \exp\{\beta'(\mathbf{z}_1 - \mathbf{z}_2)\} \qquad (15)$$

does not depend on time. Thus it is called the proportional hazards (PH) model.

Several authors (*4, 6, 10, 15-17*) have developed the likelihood analysis of the PH model from rather distinct points of view. Providing that there are no ties in the uncensored data, all derive for the ln-likelihood function of the expression

$$L(\beta) = \sum_{k=1}^{K} \{\beta'\mathbf{z}_k - \ln \sum_{j \in R(t_k)} \exp(\beta'\mathbf{z}_j)\} \qquad (16)$$

Table 6. Stratified analysis of additive maintenance therapy for childhood leukemia according to diagnostic white blood count.

White blood count (stratum)	Treated patients[a]			Controls[a]		
	N	O	E	N	O	E
0-4,999	51	25	27.08	36	18	15.92
5,000-19,999	58	38	44.64	37	18	21.16
20,000 +	43	37	39.04	43	35	32.96
Totals	152	100	110.96	116	81	70.04

[a] N = number of patients; O = observed numbers of relapses; E = expected numbers of relapses.

Table 7. Illustration of fitting of proportional hazards regression model to data on 268 children with acute leukemia.

No. of vars	Ln-Likelihood	Regression coefficients			
		log(WBC)	Age	Age^2	RX (0/1)
0	−900.84	—	—	—	—
1	−881.57	0.783 (6.367)[a]	—	—	—
3	−877.99	0.737 (5.876)	−0.186 (−2.290)	0.0145 (2.621)	—
4	−876.95	0.721 (5.709)	−0.189 (−2.316)	0.0148 (2.647)	−0.220 (−1.452)

[a] Standardized regression coefficients in parentheses.

where $R(t_k)$ is the risk set of subjects still alive and under observation just prior to t_k; $\mathbf{z}_k$ is the covariate vector for the individual who dies at t_k; and the outer summation is over all K true (uncensored) times of death. Different likelihoods arise in the case of ties.

Taking the vector of first partial derivatives of L, setting equal to $\mathbf{O}$ and solving the resulting nonlinear equations yields a maximum likelihood estimate $\hat{\beta}$ for the regression coefficients. A covariance matrix for this estimate is obtained in the usual fashion by inversion of the negative of the matrix of second partials of L. The integral

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du \tag{17}$$

defines the cumulative hazard function for the standard covariate set. Once $\hat{\beta}$ is obtained this may be estimated by

$$\hat{\Lambda}_0(t) = \sum_{t_k \leq t} (\sum_{j \in R(t_k)} \exp\{\beta'\mathbf{z}_j\})^{-1} \tag{18}$$

where the outer summation is again over true survival times t_k less than or equal to t. The corresponding estimate of the survival function

$$P_0(t) = \exp\{-\Lambda_0(t)\} \tag{19}$$

is

$$\hat{P}_0(t) = \prod_{t_k \leq t} 1 - \frac{1}{\sum_{j \in R(t_k)} \exp(\hat{\beta}' z_j)} \tag{20}$$

Notice that when $\hat{\beta} = \mathbf{0}$ this reduces to the PL estimate of Kaplan and Meier, calculated from the entire set of observations considered as one homogeneous sample.

Table 7 illustrates the computer fitting of the PH model to the data on 268 leukemic children. Four regression variables were considered: z_1 = log (WBC), z_2 = age at diagnosis (years), $z_3 = z_2^2$, and z_4 = 1 or 0, according as the patient was treated with AMD or was a control. These four variables were entered into the regression equation sequentially in order to demonstrate their effects on remission duration after adjustment for the preceding variables. A quadratic term in the age variable was required: children in the mid ranges from 2 to 6 years have a better prognosis than at either extreme. The multiplicative effect of treatment on the relapse rate is given by exp (β_4) = exp $(-0.220) = 0.80$, which is quite comparable with the approximate value 0.78 obtained from the simpler stratified analysis. The likelihood ratio test for treatment effectiveness yields a chi-square of $2(-876.95 + 877.99) = 2.09$; squaring the standardized regression coefficient gives a similar value $(-1.45)^2 = 2.11$. These are both even smaller than the value of 2.80 obtained after adjustment for WBC in three strata, so that the regression approach has in this case led to an even greater reduction in the statistical significance of the treatment comparison.

As a means of providing a graphical display of the fit of the model, the regression coefficients were used to calculate a prognostic score for each child using the formula

$$S = z_1\hat{\beta}_1 + z_2\hat{\beta}_2 + z_3\hat{\beta}_3 + z_4\hat{\beta}_4 \tag{21}$$

Each of the covariates z was first normalized by subtracting off the mean, so that a score $S = 0$ represents a "typical" patient. The scores were then used to divide the sample into four groups within each of which PL estimates of the remission duration were calculated. These are plotted in Figure 3 together with predicted remission duration curves, estimated from the model, for specified values of S. Notice that the predicted curves lie further apart than do the "observed" curves for later days, while the reverse is true for earlier days. This behavior indicates a certain lack of fit of the model, namely that the baseline covariates have more of an effect on early rates of relapse than they do on later ones. The fit could be improved by use of time-dependent covariates $\mathbf{z}(t)$ as discussed by Cox (*6*).

The PH Model in Epidemiology: Applications to RERF Data

The methodology of survival analysis, especially the PH model developed in the last section, is also useful with epidemiologic studies of risk factors for chronic disease. In this context t often represents age, and the endpoint is diagnosis of, or death from, a particular disease in a previously disease-free individual. $\lambda_0(t)$ may then be interpreted as the age-specific incidence or mortality rate for a standard covariate set, while exp $(\beta'\mathbf{z})$ represents the relative risk or rate ratio (RR) for a subject with covariates $\mathbf{z}$. These are computed from the presumed risk factors under investigation and may themselves depend on age. Application of the previously discussed methodology based on the PH model is straightforward, at least in principle. A slight modification is that the risk sets $R(t_k)$ may change with age not only due to the loss of individuals from further observation, as in clinical trials, but also from the enrollment of new subjects in the study at later ages.

Prospective epidemiologic studies involve a cohort of disease-free persons who are enrolled at

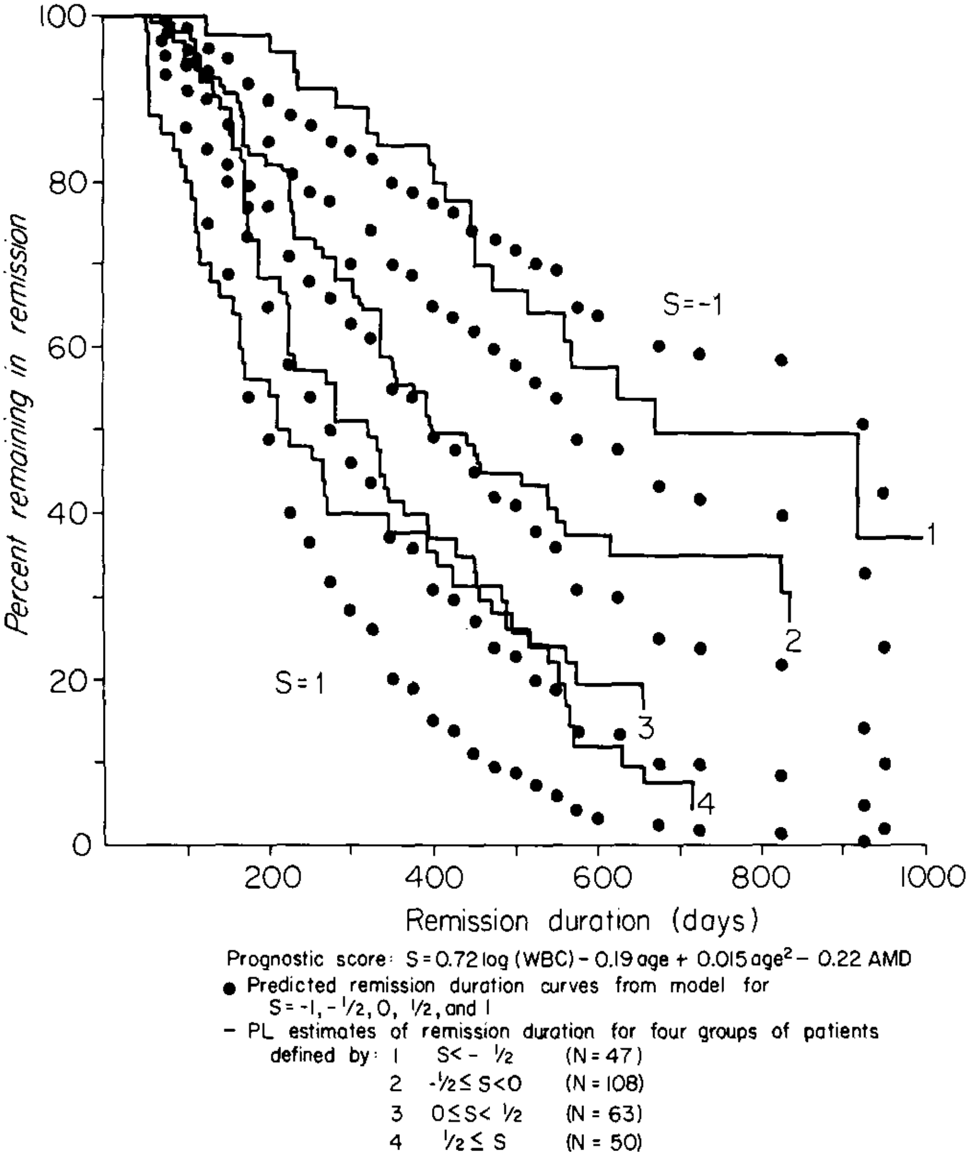

FIGURE 3. Remission duration curves for four groups of leukemic children according to their prognostic score, and predicted curves on basis of the PH model.

various ages and kept under continuous surveillance until they either develop the disease, or else are lost to further observation or die from another cause. In order to collect enough cases of rare chronic diseases, tens or even hundreds of thousands of persons may have to be followed over several years. For example, the Radiation Effects Research Foundation (RERF) has kept nearly 100,000 survivors of the Hiroshima and Nagasaki atomic bomb blasts under surveillance for nearly three decades, losing fewer than 0.1% to emigration. With such large cohorts the previously discussed iterative methods for estimation and testing of the parameters β and λ_0 are neither feasible nor necessary.

One method of dealing with large cohorts is to partition the time or age axis into intervals, say ten or twenty, and to postulate a discrete time model for the conditional probabilities of failure within each one. Suppose there are K such intervals $(0,t_1[, (t_1,t_2], \ldots, (t_{K-1}, t_K]$, and set

$$p_k(\mathbf{z}) = P[T \leq t_k | T < t_{k-1}, \mathbf{z}] \qquad (22)$$

for $k = 1, 2, \ldots, K$, where T denotes the random failure time (age at diagnosis or death), for a subject with covariates **z**. Such an individual contributes a term $p_k(\mathbf{z})$ to the likelihood if he develops the disease in the k-th interval, and a term $\{1 - p_k(\mathbf{z})\}$ if he survives it. If he dies of other causes (is censored) in the interval, Thompson (*18*) makes the sensible recommendation that he contribute a term $\sqrt{1 - p_k(\mathbf{z})}$, to reflect his survival disease-free over only a part of the interval.

Cox (*6*) suggested the use of the linear logistic model

$$\text{logit } p_k(\mathbf{z}) = \alpha_k + \beta'\mathbf{z} \qquad (23)$$

where

$$\text{logit } (p) = \ln \{p/(1 - p)\} \qquad (24)$$

This reduces to the PH model in the limit as K increases and the time or age intervals become infinitely small. The term exp $(\beta' \mathbf{z})$ in Eq. (23) represents the odds ratio of disease occurrence in each interval,

$$\frac{p_k(\mathbf{z})\{1 - p_k\,(\mathbf{0})\}}{\{1 - p_k\,(\mathbf{z})\}\, p_k(\mathbf{0})} = \exp \{\beta' \mathbf{z}\} \qquad (25)$$

rather than the ratio of instantaneous failure rates.

A similar model was one of several used by Otake (*19*) to explore the effects of radiation on cause-specific mortality using RERF data. He divided the sample into five groups according to age at the time of bomb (ATB) and six classes according to estimated total radiation doses, and considered a variety of causes of death as the endpoint. If n_{ij} denotes the number of subjects in age group i and radiation class j, while d_{ij} denotes the number of deaths due to specific cause, then his model states

$$\text{logit } E(d_{ij}/n_{ij}) = \alpha_i + \beta_j \qquad (26)$$

subject to appropriate constraints on the parameters. This is a linear logistic model for the unconditional probabilities of death over a defined period (in this case 1950-1972) and so does not explicitly account for competing causes of death or differential losses to further observation which may be taking place in the 30 age × dose cells.

Cox's model based on the conditional probabilities would further divide the follow-up period into K intervals, say 1950-1954, 1955-1959, etc. If n_{ijk} then denotes the number of subjects who were age i ATB, received radiation dose j, and who remained alive at the midpoint of the k-th interval, the model is

$$\text{logit } E(d_{ijk}/n_{ijk}) = \alpha_i + \beta_j + \gamma_k \qquad (27)$$

One might include also an interaction term $(\alpha\gamma)_{ik}$ so as to allow age and calendar time to have essentially arbitrary effects on disease incidence. The numbers of deaths d_{ijk} may be formally considered to have independent binomial distributions conditionally on the n_{ijk} (*4*).

A different approach to this problem, termed by Mantel a "synthetic retrospective study," is to draw a small sample of disease-free "controls" from each of the risk sets $R(t_k)$ for comparison with the diseased cases as they arise. The very same theory and methods may be applied also to actual case/control studies in which cases are ascertained as they occur, for example through a population based disease register. The controls, instead of being obtained from a computer file, are samples from the population in which the case arose. They may be patients with different diagnoses in the same hospital, chosen from the same family or neighborhood, or simply sampled at random from the population.

While it is impossible to estimate the incidence rates $\lambda_0(t)$ without studying the full cohort, the PH model assumed for the prospective study implies a probability structure for the sample cases and controls which may be used to estimate the parameter β describing the RR associated with the covariates (*20*). Specifically, suppose $m = m(t)$ cases having covariate vectors $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are diagnosed (or die) with the particular disease at age t. Suppose also that n disease-free controls with covariates $\mathbf{z}_{m+1}, \ldots \mathbf{z}_{m+n}$ are drawn at random from the risk set. The analysis is performed conditionally on fixed values for m, n and the combined set of $n + m$ observed covariate vectors. Using Cox's linear logistic model for the conditional failure probabilities, the likelihood that m individuals with indices $l_1, \ldots, l_m$ are cases and the remainder are controls, given that they all survive to age t, is

$$\prod_{i=1}^{m} \exp \{(\alpha + \beta' \mathbf{z}_{l(i)} \prod_{j=1}^{m+n} (1 + \exp \{\alpha + \beta'\mathbf{z}_j\})^{-1} \qquad (28)$$

where $\alpha = \alpha(t)$. Consequently the conditional probability that the first m **z**'s correspond to cases (as observed), and the remainder to controls, is

$$\frac{\exp \{(\beta'\mathbf{s})\}}{\sum_{l\epsilon R(m,n)} \exp\, (\beta'\mathbf{s}_l)}, \qquad (29)$$

where $l = (l_1, \ldots, l_m)$ ranges over the set $R(m,n)$ of $\binom{m+n}{m}$ subsets of size m from $\{1, 2, \ldots, m + n\}$, $\mathbf{s} = \mathbf{z}_1 + \ldots + \mathbf{z}_m$, and $\mathbf{s}_l = \mathbf{z}_{l1} + \ldots + \mathbf{z}_{lm}$.

If m_k cases and n_k controls are sampled at age t_k, the conditional likelihood function for β may be written

$$\prod_{k=1}^{K} (\exp\{\beta' s_k\}/ \sum_{l \epsilon R(m_k,n_k)} \exp\{\beta' s_l\}) \tag{30}$$

This expression is formally identical to that given earlier for the PH model in survival analysis. However now the risk sets $R(t_k)$, instead of containing everyone still alive and disease-free at t_k, are replaced by the much smaller sets of $m_k + n_k$ subjects actually sampled for the retrospective analysis.

In practice, the sets of $m + n$ cases and controls may be matched on other variables **x** besides age, for example on risk factors already known to be associated with the disease under investigation. The PH model may then be generalized to

$$\lambda(t|\mathbf{z},\mathbf{x}) = \exp\{\beta'\mathbf{z}\}\, \lambda_0(t|\mathbf{x}) \tag{31}$$

which allows the underlying incidence function λ_0 to depend in an arbitrary way on **x** as well as t. Interactions between the matching variables and the risk factors continue to be modelled in **z**. Retrospective sampling carried out within risk sets having similar **x** and t values leads to the same conditional likelihood given above (*20*).

A special case of the conditional likelihood occurs when each case is individually matched to exactly R controls, a situation found often in practice. Suppose there are K such sets and denote by $\mathbf{z}_{0k}$ the covariate vector for the case and by $\mathbf{z}_{1k}, \ldots, \mathbf{z}_{Rk}$ the covariates of the controls in the k-th set, $k = 1, \ldots, K$. The conditional likelihood may then be written

$$\prod_{k=1}^{K} \frac{1}{1 + \sum_{j=1}^{R} \exp\{\beta' (\mathbf{z}_{jk} - \mathbf{z}_{0k})\}}. \tag{32}$$

This approach was taken with an RERF data file consisting of records on 7078 males who were 50+ years old ATB. Thirteen deaths from leukemia occurred (Table 8). For each of these a search was made to determine the risk sets of potential controls who had the same age ATB (exact year), were residing in the same city (Hiroshima or Nagasaki), and who were alive at the end of the same year in which the case died. These ranged in size from 30 to 508 individuals.

In order to illustrate the application of the case/control methodology, a series of 1, 2, 5, 10, and 20 controls was drawn at random from each of the risk sets. Two covariates were computed from the estimated radiation dose (in rads) for each subject: $z_1 = \ln(\text{rads} + 1)$ and $z_2 = \sqrt{\text{rads}}$. The choice of ln (rads + 1) as a covariate in the PH model, which implies there is a linear increase in ln RR with ln (rads + 1), was based on two facts: the highly skewed distribution of radiation doses; and Figure 2 of Otake (*19*), which shows a nonlinear increase in ln RR with dose after about 200 rad. The alternate transformation $\sqrt{\text{rads}}$ was used for comparison.

Results of the matched analyses based on the dif-

Table 8. Summary data on records used for matched case-control analysis of RERF files.

Case	ID	City	Age ATB	Year of death	Dose, rad	Risk set Size	Risk set Dose[1a]
1	252750[b]	H	56	1957	0	303	2
2	278624[b]	H	56	1957	14	303	1
3	090238	N	50	1967	325	73	4
4	091006	N	55	1959	163	79	2
	254898	H	60	1953	0	251	4
6	257863	H	51	1953	24	508	2
7	275949	H	51	1958	238	445	2
8	400967	H	57	1954	574	345	2
9	401117	H	63	1958	867	116	1
10	434199	H	66	1956	8	99	2
11	437743	H	52	1963	38	293	3
12	468806	H	50	1960	NIC[c]	419	2
13	083922	N	50	1976	13	30	6

[a] Geometric mean dose for 30 controls sampled at random from risk set.

[b] Note that these two cases have identical risk sets as they lived in the same city and had the same years of birth and death.

[c] Not in city ATB (0 dose).

Table 9. Results of case-control analysis of radiation exposures and leukemogenesis among males age 50+ ATB.

Covariate	Number of controls in analysis	Coefficients		RR estimate for dose			Ln-Lik
		$\hat{\beta}$	S.E.	10	50	200	
$z = \ln(\text{rads} + 1)$	1	0.284	0.201	1.98	3.06	4.51	−7.78
	2	0.434	0.205	2.83	5.50	9.97	−11.08
	5	0.482	0.163	3.18	6.66	12.91	−17.69
	10	0.415	0.134	2.71	5.12	9.05	−25.88
	20	0.475	0.131	3.12	6.48	12.43	−32.75
$z = \sqrt{\text{rads}}$	1	0.049	0.046	1.17	1.42	2.02	−8.34
	2	0.085	0.049	1.31	1.83	3.35	−12.22
	5	0.117	0.044	1.45	2.28	5.20	−18.89
	10	0.095	0.033	1.35	1.96	3.83	−27.18
	20	0.077	0.029	1.28	1.73	2.98	−35.90

Table 10. Comparison of matched and unmatched analyses of RERF files: twenty controls per case.

Covariate		Regression coefficients ± S.E.			RR estimate for dose		
		Const	z	z^2	10	50	200
$z = \ln(\text{rads} + 1)$	Matched	—	0.475 ± 0.131	—	3.12	6.48	12.43
	Unmatched	−3.971 ± 0.484	0.453 ± 0.123	—	2.96	5.93	11.04
	Unmatched	−4.105 ± 0.583	0.626 ± 0.386	− 0.027 ± 0.056	3.85	7.77	13.08
Test for quadratic effect: $\chi_1^2 = 0.23$							
$z = \sqrt{\text{rads}}$	Matched	—	0.077 ± 0.029	—	1.28	.73	2.98
	Unmatched	−3.332 ± 0.337	0.072 ± 0.027	—	1.18	1.67	2.77
	Unmatched	−3.877 ± 0.470	0.255 ± 0.100	−0.005 ± 0.004	2.20	4.71	13.33
Test for quadratic effect $\chi_1^2 = 7.34$							

ferent numbers of controls are shown in Table 9. Note the decrease in the estimated standard errors as additional controls are used; however the gain afforded by using twenty rather than ten controls is not great. The regression coefficients of 0.4–0.5 for ln (rads + 1) indicate that, roughly speaking, the risk of leukemia increases by about 5% for every 10% increase in radiation dose.

The disparity between the relative risks fitted under the model $z = \ln(\text{rads} + 1)$ vs. those fitted by $z = \sqrt{\text{rads}}$ is quite noticeable, especially when one recalls that most doses fall in the 0 – 200 range. In order to try to understand better why this might occur, unmatched logistic regression analyses were carried out on the data consisting of all cases and the sets of twenty controls. The estimated slopes and standard errors were similar to those obtained with the matched analysis (Table 10). However the intercepts, corresponding to the estimated log odds of leukemia in the sample at a dose of zero rads, differed markedly between the two models: $\alpha = -3.97$ for $z = \ln(\text{rads} + 1)$ vs. $\alpha = -3.33$ for $z = \sqrt{\text{rads}}$. Differences in the estimated relative risk between the two models were thus explained largely by the differences in the absolute risks estimated for the baseline value of the covariate.

A potential hazard of the regression modelling of relative risks is its sensitivity to the choice of scale on which the covariate is measured. Goodness of fit of the model to the data is essential for proper interpretation, and should be explored thoroughly. When both linear and quadratic covariate terms were fitted with the models above, for example, the agreement between the estimated values for α improved substantially. Moreover the fact that the quadratic term was highly significant for the $\sqrt{\text{rads}}$ model, and not at all significant for the ln (rads + 1) model, showed that the latter gave much better fit to these data (Table 10).

Further Reading

Much of the above material is presented in greater detail in a review article (*10*) on the PH model and its applications to survival data. Some additional applications of this model to epidemiologic studies are given by Breslow et al. (*21, 22*). Peto et al. (*23*) present a thorough discussion of its use in the design and analysis of clinical trials.

A computer program for calculating the PL estimate and all the test statistics presented in sections 2-5 above is available (*24*).

Several authors have pointed out that the *W* scores defined do not lead to the most efficient generalization of Wilcoxon's test to censored data. They all propose essentially the same statistic as an alternate generalization (*7, 25, 26*).

A comparison of the efficiencies of the test statistics using Monte Carlo techniques is made by Lee et al. (*27*). Efron (*17*) discusses the efficiency of the likelihood function used with the PH model from a more abstract viewpoint; see also Kalbfleisch (*28*).

Additional extensions of the PH regression model for use with grouped or heavily tied data are discussed by Cox (*6*), Kalbfleisch and Prentice (*15*), Thompson (*18*), and Prentice and Gloeckler (*29*).

This work was supported in part by Grant CA-11722 from the U.S. National Institutes of Health.

REFERENCES

1. Heyn, R. M., Holland, R., Newton, W. A., Jr., Tefft, M., Breslow, N., and Hartman, J. R. The role of combined chemotherapy in the treatment of rhabdomyosarcoma in children. Cancer 34: 2128 (1974).
2. Kaplan, E. L., and Meier, P. Non-parametric estimation from incomplete observation. J. Am. Statist. Assoc. 53: 457 (1958).
3. Breslow, N., and Crowley, J. A large sample study of the life table and product limit estimates under random censorship. Ann. Statist. 2: 437 (1974).
4. Cox, D. R. Partial likelihood. Biometrika 62: 269 (1975).
5. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chem. Repts. 50: 163 (1966).
6. Cox, D. R. Regression models and life tables (with discussion) J. Roy. Statist. Soc. B34: 187 (1972).
7. Peto, R., and Peto, J. Asymptotically efficient rank invariant test procedures. J. Roy. Statist. Soc. A135: 185 (1972).
8. Petro, R., and Pike, M. C. Conservatism of the approximation $\Sigma (O - E)^2/E$ in the log rank test for survival data or tumor incidence data. Biometrics 29: 579 (1973).
9. Crowley, J., and Breslow, N. Remarks on the conservatism of $\Sigma(\mathrm{O}-\mathrm{E})^2/\mathrm{E}$ in survival data. Biometrics 31: 957 (1975).
10. Breslow, N. Analysis of survival data under the proportional hazard model. Intern. Statist. Rev. 43: 43 (1975).
11. Gehan, E. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 52: 203 (1965).
12. Breslow, N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 57: 579 (1970).
13. Tarone, R. E. Tests for trend in life table analysis. Biometrika 62: 679 (1975).
14. Miller, D., Sonley, M., Karon, M., Breslow, N., and Hammond, D. Additive therapy in the maintenance of remission in acute lymphoblastic leukemia of childhood: The effect of the initial leukocyte count. Cancer 34: 508 (1974).
15. Kalbfleisch, J. D., and Prentice, R. L. Marginal likelihoods based on Cox's regression and life model. Biometrika 60: 267 (1973).
16. Breslow, N. Covariance analysis of censored survival data. Biometrics 30: 89 (1974).
17. Efron, B. The efficiency of Cox's likelihood function for censored data. J. Am. Statist. Assoc. 72: 557 (1977).
18. Thompson, W. A. On the treatment of grouped observations in life studies. Biometrics 33: 463 (1977).
19. Otake, M. Radiation effects on cancer mortality among A-bomb survivors, 1950-72: Comparison of some statistical models and analyses based on the additive logit model. J. Radiation Res. 17: 262 (1976).
20. Prentice, R. L., and Breslow, N. E. Retrospective studies and failure time models. Biometrika 65: 153 (1978).
21. Breslow, N. The proportional hazards model: applications in epidemiology. Comm. Statist. Theor. Meth. 47: 315 (1978).
22. Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. Estimation of multiple relative risk functions in matched case-control studies. Amer. J. Epidemiol. 108: 299 (1978).
23. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and Design. Brit. J. Cancer 34: 585 (1976); II. Analysis and examples. Brit. J. Cancer 35: 1 (1977).
24. Thomas, D. G., Breslow, N., and Gart, J. Trend and homogeneity analyses of proportions and life table data. Computers Biomed. Res. 10: 373 (1977).
25. Efron, B. The two sample problem with censored data. Proc. Fifth Berkeley Symp. 4: 831 (1967).
26. Prentice, R. L. Linear rank tests with right censored data. Biometrika 65: 167 (1978).
27. Lee, E. T., Desu, M. M., and Gehan, E. A Monte Carlo study of the power of some two sample tests. Biometrika 62: 425 (1975).
28. Kalbfleisch, J. D. Some efficiency computations for survival distributions. Biometrika 61: 31 (1974).
29. Prentice, R. L. and Gloeckler, L. Regression analysis of grouped survival data with application to the study of breast cancer survival times. Biometrics 34: 57 (1978).